Neuroscience 228 (2013) 361–370

WHAT DO I DO NOW? AN ELECTROENCEPHALOGRAPHIC INVESTIGATION OF THE EXPLORE/EXPLOIT DILEMMA

C. D. HASSALL,* K. HOLLAND AND O. E. KRIGOLSON

Department of Psychology and Neuroscience, Dalhousie University, Halifax, Nova Scotia, Canada B3H 4R2

Abstract—To maximize reward, we are faced with the dilemma of having to balance the exploration of new response options and the exploitation of previous choices. Here, we sought to determine if the event-related brain potential (ERP) in the P300 time range is sensitive to decisions to explore or exploit within the context of a sequential risk-taking task. Specifically, the task we used required participants to continually explore their options—whether they should ''push their luck'' and keep gambling or ''take the money and run'' and collect their winnings. Our behavioral analysis yielded two distinct distributions of response times: a larger group of short-decision times and a smaller group of long-decision times. Interestingly, these data suggest that participants adopted one of two modes of control on any given trial: a mode where they quickly decided to keep gambling (i.e. exploit), and a mode where they deliberated whether to the take the money they had already won or continue gambling (i.e. explore). Importantly, we found that the amplitude of the ERP in the P300 time range was larger for explorative decisions than for exploitative decisions and, further, was correlated with decision time. Our results are consistent with a recent theoretical account that links changes in ERP amplitude in the P300 time range with phasic activity of the locus coeruleus–norepinephrine system and decisions to engage in exploratory behavior.

Key words: P300, exploration/exploitation tradeoff, decision making, reinforcement learning, ERP, learning.

INTRODUCTION

In Mill's Utilitarianism (1863/2008), he argued that humans have an inherent desire to maximize utility. As such, the decisions that we make on a day-to-day and moment-to-moment basis typically reflect a desire to maximize the reward. However, as Dennett (1986) and others have pointed out, calculating the utility of decisions in the real world can be challenging because the potential consequences of our actions are not always known. Even if utility calculations are restricted to the near future, complex or novel situations may arise that require exploring options with unknown consequences. Exploration is inherently risky but necessary in order to assess new response options or reassess old ones. The knowledge gained through exploration can later be exploited to improve subsequent decisions, and thus yield even greater increases in utility. However, one cannot always engage in exploratory behavior. Rather, one must balance exploratory behavior with exploitation—selecting the most rewarding response option as much as possible. Therefore, an optimal decision strategy for maximizing utility would entail utilizing an exploitative mode of control most of the time with occasional instances of exploratory behavior.

Experimentally, decisions to explore or exploit can be studied in tasks such as the Balloon Analog Risk Task (BART: Lejuez et al., 2002). During performance of the BART, participants must continually explore their options—either take the money they have already earned or continue gambling. The key manipulation of the BART is that, for each pump of the balloon (gamble), the amount of money earned increases along with the probability of losing all earned money. This manipulation makes each gamble increasingly risky. Thus, there is an optimal response in the BART (i.e. total number of balloon pumps) that is based on the risk and reward structure of the task (Lejuez et al., 2002), and as such, to maximize reward, participants need to explore in order to determine the optimal number of balloon pumps. Computational models of the BART suggest that people make a risk assessment prior to each pump: a decision to continue pumping or collect their accumulated reward (Wallsten et al., 2005). The Wallsten et al. (2005) model's predictions were recently corroborated by Wershbale and Pleskac (2010) who observed two distinct distributions of response times in human BART performance. Specifically, they observed that people generally made automatic, rapid responses in the BART, but occasionally paused to assess whether or not they should continue gambling. Wershbale and Pleskac (2010) hypothesized that these pauses represent the assessments predicted by earlier modeling work (Wallsten et al., 2005; Pleskac, 2008). Interestingly, the number of assessments that

*Corresponding author. Address: Department of Psychology and Neuroscience, Dalhousie University, P.O. Box 15000, Halifax, Nova Scotia, Canada B3H 4R2. Tel: +1-902-494-2923; fax: +1-902-494-6585.

E-mail address: cameron.hassall@dal.ca (C. D. Hassall).

Abbreviations: BART, Balloon Analog Risk Task; BSR, Bayesian sequential risk-taking model; EEG, electroencephalographic; ERP, event-related brain potential; fMRI, functional magnetic resonance imaging; LC–NE, locus coeruleus–norepinephrine; PFC, prefrontal cortex; sLORETA, standardized low-resolution brain electromagnetic tomography.

0306-4522/12 $36.00
http://dx.doi.org/10.1016/j.neuroscience.2012.10.040

participants made during the BART decreased over time. Importantly, this change in assessment rate is consistent with theoretical models of the exploration/exploitation dilemma. Early in learning, people need to explore more often in order to determine the reward structure of a task (e.g., the optimal number of pumps in the BART). However, once the reward structure is known, people exploit more frequently. With all of this in mind, Wershbale and Pleskac (2010) likened fast BART responses to exploitation and slower responses to exploration.

Research examining the neural basis of decisions to explore or exploit is limited (see Cohen et al., 2007 for a review). In one recent study, Cavanagh et al. (2011) suggested increased frontal theta-band oscillation as a possible neural marker of uncertainty-driven exploration. Specifically, Cavanagh and colleagues (2011) observed a correlation between medial–frontal theta power and the parameters of their reinforcement-learning model during exploration in a decision-making task. From their results, Cavanagh et al. (2011) hypothesized that midbrain regions were responsible for exploitation but that frontal brain regions took control when deciding to explore in uncertain situations. The Cavanagh et al. (2011) hypothesis is consistent with an earlier functional magnetic resonance imaging (fMRI) study that showed enhanced frontal brain activity during exploratory decisions in a four-armed bandit task (Daw et al., 2006). Cavanagh and colleagues' (2011) hypothesis is also consistent with work by Frank et al. (2009) that associated a prefrontal cortex (PFC) dopamine gene (COMT) with exploratory decisions. In particular, Frank et al. (2009) showed an effect of COMT gene dose (which they defined as the amount of methionine-encoding or *met* allele present) on uncertainty-driven exploration. The presence of the met allele is linked to increased PFC dopamine levels compared to the presence of the valine-encoding or *val* allele. Although Frank et al. (2009) were uncertain about the exact role of COMT in exploratory behavior, they suggested that the observed and known effects of the met allele implicate the PFC as the controller of uncertainty-driven exploration. Taken together, these studies suggest that switching from an exploitative to an explorative mode of control involves the intervention of frontal cognitive systems over midbrain lower-level reward-processing systems (see Mars et al., 2011, for more examples of cognitive control).

Currently, there are no definitive electroencephalographic (EEG) correlates differentiating decisions to explore or exploit. Having said that, there are good reasons to hypothesize that the event-related brain potential (ERP) in the time range of the P300 may be sensitive to this distinction. The P300 is a high-amplitude, positive ERP component with peak latency 300–500 ms following the presentation of a stimulus (Sutton et al., 1965) that has been associated with several different cognitive functions (Polich, 2007). One influential account—the context-updating hypothesis—states that the P300 reflects the updating of an internal model of the probabilistic structure of the world (Donchin, 1981; Donchin and Coles, 1988). Donchin's (1981) account arose out of earlier observations that the P300 is sensitive to stimulus frequency (Duncan-Johnson and Donchin, 1977). Consistent with the context-updating hypothesis, Nieuwenhuis et al. (2005) recently suggested that ERP changes in the P300 time range reflect the locus coeruleus–norepinephrine (LC–NE) system's response to internal decision-making processes regarding task-relevant stimuli (Aston-Jones and Cohen, 2005; Nieuwenhuis, 2011; also see Pineda et al., 1989, for early work linking the LC and the P300). The LC contains noradrenergic neurons and provides the only source of NE to the hippocampus and neocortex (Berridge and Waterhouse, 2003). Increases in LC activity, and the associated rise in NE, are linked to increased exploratory behavior in monkeys (Aston-Jones and Bloom, 1981; Usher et al., 1999; Aston-Jones and Cohen, 2005; modeled by McClure et al. (2006)). Importantly, a series of lesion, psychopharmacological, and EEG studies support the link between an ERP difference in the P300 time range and phasic changes in the activity of the LC–NE system (see Nieuwenhuis et al., 2005, for a review). Thus, given the link between the LC–NE system and exploration, and the link between the LC–NE system and the P300, it stands to reason that the amplitude of the ERP in the P300 time range may differentiate decisions to explore or exploit.

Our main purpose here was to determine whether or not ERP amplitude in the P300 time range would be sensitive to decisions to explore or exploit. To accomplish this, we had participants perform a modified version of the BART while EEG data were recorded. In terms of behavior, we expected to observe a similar distribution of response times as Wershbale and Pleskac (2010). In particular, we expected to see two distinct distributions of response times: one distribution of fast responses indicative of exploitation, and a second distribution of slow responses indicative of exploration. Importantly, we predicted that the amplitude of the ERP in the P300 time range preceding decisions to explore would be greater than the ERP amplitude in the same time range preceding decisions to exploit—a prediction derived from Nieuwenhuis and colleagues' (2005) hypothesis that ERP modulation in the P300 time range is driven by phasic changes in LC–NE activity linked to internal decision-making processes.

There is a growing body of evidence that the amplitude of the P300 is also modulated by reward magnitude (Yeung and Sanfey, 2004; Hajcak et al., 2005; Bellebaum and Daum, 2008; Wu and Zhou, 2009). The P300's sensitivity to reward magnitude is of particular importance here because the purpose of exploration is to specify or update values associated with actions, and the purpose of exploitation is to take advantage of current value assessments (Sutton and Barto, 1998). As such, we also hypothesized that the amplitude of the P300 elicited by balloon bursts would scale with the magnitude of the amount of lost reward, reflecting an update of participants' model of the probabilistic reward structure of the task.